JOURNAL OF MULTIVARIATE ANALYSIS 15, 63–72 (1984)

Strong Consistency of Nearest Neighbor Regression Function Estimators

PHILIP E. CHENG*

University of Houston

Communicated by the Editors

For a well-known class of nonparametric regression function estimators of nearest neighbor type the uniform measure of deviation from the estimators to the true regression function is studied. Under weak regularity conditions it is shown that the estimators are uniformly consistent with probability one and the corresponding rate of convergence is near-optimal.

1. INTRODUCTION

Consider nonparametric estimation of the regression function $R(x) = E(Y \mid X = x)$ given a random sample $(X_1, Y_1),..., (X_n, Y_n)$ from an unknown joint distribution function. The basic idea of nonparametric estimation by nearest neighbor rules was suggested by Fix and Hodges [7] and formalized by Royall [10]. A brief review of the theoretic development in recent years is worth mentioning. Results on the pointwise mean square consistency and the asymptotic normality are found in [10], global mean L_p-consistency is given in [12] and [6], weak convergence in the form of one-dimensional stochastic process is given in [3] and uniform strong consistency is established in [4]. The objective of this paper is to show that the uniform strong consistency is valid under regularity conditions weaker than those stated in [4] and the associated near-optimal rate of convergence is obtained.

Let $(X_1, Y_1),..., (X_n, Y_n)$ be a random sample from the joint distribution of (X, Y) where X is R^d-valued and Y is real-valued. For each x in R^d, order the pairs (X_i, Y_i), $i = 1,..., n$, according to the nondecreasing distances $\rho(X_i, x)$, $i = 1,..., n$, and obtain $(X_{\sigma(1;x)}, Y_{\sigma(1;x)}),..., (X_{\sigma(n;x)}, Y_{\sigma(n;x)})$ where

Received January 18, 1982; revised July 7, 1983.

AMS 1980 Subject classification: Primary 62G05.

Key words and phrases: Strong consistency, regression function, near neighbour rules.

* P. E. Cheng is currently a visiting research associate at the Institute of Statistics, Academia Sinica, Taipei, Taiwan. The research was partially supported by the Army, Navy and Air Force under Office Naval Research Contract N00014-79-C-0801.

0047-259X/84 $3.00

683/15/1-5

$\sigma(i; x)$, $i = 1,..., n$, is the corresponding permutation of the set $\{1,..., n\}$. Through this paper ρ is either the usual Euclidean metric or the maximum component norm on R^d. Following the definitions considered in [4, 10, 12] we denote a nearest neighbor estimator of $R(x)$ by

$$T_n(x) = \sum_{i=1}^{n} C_{ni} Y_{\sigma(i;x)} \tag{1.1}$$

where for each n the weights C_{ni}, $i = 1,..., n$, are selected to satisfy the conditions

(i) $\sum_{i=1}^{n} C_{ni} = 1$,

(ii) $\sum_{i=k+1}^{n} C_{ni} \to 0$, where $k = [k(n)]$ ($[t]$ denotes the smallest integer not less than t throughout this paper) is a nondecreasing sequence of integers such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$,

(iii) $\max_i C_{ni} \to 0$ as $n \to \infty$. (1.2)

In case ties of the distances $\rho(x, X_i)$ occur, we may let the permutation $\sigma(\cdot, x)$ be arbitrarily determined within each tied subset of the X_i's and assign equal weights to the corresponding Y_i's. A special case that has been paid much attention is the k-nearest neighbor estimators whose weights satisfy the conditions

(i) $\sum_{i=1}^{n} C_{ni} = 1$,

(ii) $C_{ni} = 0$ for $i > k$, $k = [k(n)]$, $k \to \infty$ and $k/n \to 0$, (1.3)

(iii) $\max_i C_{ni} \leqslant C/k$ for a positive constant $C \geqslant 1$.

Examples of the k-nearest neighbor weights (see [12]) include the uniform weights with $C_{ni} = 1/k$, $i = 1,..., k$, and the triangular and the quadratic weights. A simplified version of the k-nearest neighbor estimator (see [4]) is $\hat{T}_n(x) = T_n(X_{\sigma(1;x)})$ which greatly facilitates computation when n is large. The large sample performance of the estimator $\hat{T}_n$ is shown to be comparable to that of T_n in [4]. It will be indicated in Corollary 2 that the same is true regarding the rates of convergence.

In this paper, uniform strong convergence of T_n to R, i.e., $\|T_n - R\|_B = \sup_{x \in B} |T_n(x) - R(x)| \to 0$ with probability 1, is obtained under regularity conditions weaker than those required in [4], namely, we assume that the noise about the true regression function has finite variance. The associated convergence rate, $n^{1/(2+d)}(\beta_n \log n)^{-1} \|T_n - R\|_A \to 0$ with probability 1, comparable to the result for the kernel regression estimator (see [1]), is also obtained under milder restrictions. In general, asymptotic results which are valid for estimators with weights (1.3) are also true for those with weights (1.2) provided that the quantity $\sum_{i=k+1}^{n} C_{ni}$ approaches zero sufficiently fast (see for examples Theorem 2 in [12] and Theorem 4.1 in [6]). For ease of exposition, our results will be stated and proved only for the k-nearest neighbor estimators although extensions to the wider class of estimators with weights (1.2) are straightforward.

2. Results

Let $(X_1, Y_1),..., (X_n, Y_n)$ be a random sample from the joint distribution of (X, Y) on $R^d \times R^1$. Let μ be the Borel probability measure associated with the distribution of X and B be the support of μ. Throughout this paper we shall assume that B is a bounded hence compact subset of R^d. The symbol $S(x; r)$ denotes the ρ-metric ball with radius r centered at x. For each x in B, let $R_k(x) = \rho(X_{\sigma(k;x)}, x)$ be the distance from the point x to the kth-nearest X_i of the sample. To abbreviate notations, we shall write "w.p.l" for "with probability 1" and suppress the argument "x" writing R_k and $\sigma(i)$, $i = 1,..., n$. Following a definition in [4], we say that the L_t condition, $t > 0$, holds if there exists a finite positive constant D_t such that $\sup_{x\in B} E(|Y - R(x)|^t \,|X = x) \leqslant D_t$. The following regularity conditions will be assumed throughout this paper.

(i) R is continuous on B,

(ii) the L_2 condition holds, i.e., $\mathrm{Var}(Y|X)$ is a bounded random variable.

(2.1)

Notice that (2.1) implies that R is bounded on B and $EY^2 < \infty$. We now state our first result for the k-nearest neighbor estimators T_n.

Theorem 1. *Assume conditions* (1.3) *and* (2.1). *Select k such that* $k/\sqrt{n} \log n \to \infty$. *Then*

$$\|T_n - R\|_B \to 0 \qquad w.p.\ 1.$$

Corollary 1. *Under the conditions of Theorem* 1,

$$\|\hat{T}_n - R\|_B \to 0 \qquad w.p.\ 1.$$

We remark that the conclusion of Theorem 1 was obtained in [4] under more restrictive L_t conditions with $t > 3 + d$ for the Euclidean metric and with $t > 2d + 2$ for the maximum component metric. It was also obtained in [6] under the assumption that Y is a bounded random variable. Note that our choice of k is more restrictive than those required in [4] and [6]. This appears to be the way it should be. The L_2 condition is theoretically convenient and is utilized for obtaining convergence rates for other nonparametric regression function estimators (see [1, 11, 13]).

Our second result provides a rate of the uniform convergence of Theorem 1. Conditions stronger than (2.1) will be imposed.

(i) R is uniformly locally Lipschitz of order 1 on B, i.e., there exist positive constants α and δ such that $|R(x)-R(y)| \leqslant \alpha\rho(x,y)$ if $x, y \in B$ and $\rho(x,y) \leqslant \delta$,

(ii) $E\,|Y|^{2+d} < \infty$,

(iii) L_2 condition holds,

(iv) μ is absolutely continuous with a continuous derivative f relative to the Borel measure on R^d. (2.2)

THEOREM 2. *Assume* (1.3) *and* (2.2) *and select* $k = [Cn^{2/(2+d)}]$. *Then, for any bounded set A where f is bounded away from zero,*

$$n^{1/(2+d)}(\beta_n \log n)^{-1}\,\|T_n - R\|_A \to 0 \qquad w.p.\ 1,$$

where $\beta_n \to \infty$ *arbitrarily.*

COROLLARY 2. *Under the conditions of Theorem* 2,

$$n^{1/(2+d)}(\beta_n \log n)^{-1}\,\|\hat{T}_n - R\|_A \to 0 \qquad w.p.\ 1.$$

Compared to the optimal rates of convergence in probability (see [13]) for nonparametric estimators, the rates given in Theorem 2 are near-optimal aside from the factor $(\beta_n \log n)$ where β_n diverges to infinity at any slow rate. The conditions and rates of Theorem 2 are comparable to those for the kernel estimators (see [1]) except that we do not impose differentiability condition on the density function f. Furthermore, (2.2)(iv), continuity of the density function f, may be relaxed requiring simply that

(iv′) f be bounded away from zero over some open set containing A.

In this case, a result slightly weaker than Theorem 2 holds.

THEOREM 3. *Assume the conditions of Theorem* 2 *but replacing* (2.2)(iv) *with* (iv′) *above. Then for any* $\varepsilon > 0$,

$$n^{(1-\varepsilon)/(2+d)}\,\|T_n - R\|_A \to 0 \qquad w.p.\ 1.$$

COROLLARY 3. *Under the conditions of Theorem* 3,

$$n^{(1-\varepsilon)/(2+d)}\,\|\hat{T}_n - R\|_A \to 0 \qquad w.p.\ 1.$$

3. Proofs

We shall begin with a few preliminary lemmas. Firstly, a fact about the conditional distribution of the $Y_{\sigma(i)}$'s given the $X_{\sigma(i)}$'s is cited from [10].

LEMMA 1. *If $(X_1, Y_1),..., (X_n, Y_n)$ are independent, then, for Borel subsets $B_1,..., B_n$ of R^1 and $\sigma(\cdot)$ being a function of the X_i's,*

$$P[Y_{\sigma(i)} \in B_i, i = 1,..., n \mid X_{\sigma(i)} = x_i, i = 1,..., n]$$
$$= \prod_{i=1}^{n} P[Y_{\sigma(i)} \in B_i \mid X_{\sigma(i)} = x_i] \qquad w.p.\ 1.$$

In view of Lemma 1, we define

$$V_n(x) = E(T_n(x) \mid X_{\sigma(i)}, i = 1,..., n)$$

and note that $V_n(x) = \sum_{i=1}^{n} C_{ni} R(X_{\sigma(i)})$ w.p. 1. Next, a result from [4] is stated as follows.

LEMMMA 2. *Assume conditions* (1.2) *and* (2.1)(i). *Then*

$$\| V_n - R \|_B \to 0 \qquad and \qquad \| R_k \|_B \to 0 \qquad w.p.\ 1. \tag{3.1}$$

For proof of Theorem 1, we will use a truncation argument. Define for each positive integer n

$$\bar{T}_n(x) = \sum_{i=1}^{n} C_{ni} \bar{Y}_{\sigma(i)},$$

where

$$\bar{Y}_j = Y_j I[|Y_j| \leqslant n^{1/2}]$$

for $1 \leqslant j \leqslant n$ and $I[\cdot]$ is the usual indicator function. Set

$$\bar{V}_n(x) = E(\bar{T}_n(x) \mid X_{\sigma(i)}, i = 1,..., n)$$
$$= \sum_{i=1}^{n} C_{ni} E(\bar{Y}_{\sigma(i)} \mid X_{\sigma(i)}) \qquad \text{w.p. } 1.$$

LEMMA 3. *Assume* (1.3) *and* $EY^2 < \infty$. *Then*

$$\| T_n - \bar{T}_n \|_B \to 0 \qquad w.p.\ 1. \tag{3.2}$$

Proof. Given that $EY^2 < \infty$, it is well known that

$$P(|Y_j| > j^{1/2} \text{ infinitely many } j) = 0.$$

Thus, for some full set Ω ($P(\Omega)=1$), there exists for each $\omega \in \Omega$ a finite set of positive integers $J_\omega = \{j\colon Y_j(\omega) > j^{1/2}\}$. Hence, for each $\omega \in \Omega$ and each positive integer n, $Y_{\sigma(i)}(\omega) = \bar{Y}_{\sigma(i)}(\omega)$ for all i except for some $\sigma(i) \in J_\omega$. Since J_ω is finite and independent of n and x, it follows that for each $\omega \in \Omega$

$$\begin{aligned}\|T_n(x,\omega) - \bar{T}_n(x,\omega)\|_B &\leqslant \sum_{i=1}^{k} C_{ni} \|Y_{\sigma(i)}(\omega) - \bar{Y}_{\sigma(i)}(\omega)\|_B \\ &\leqslant \frac{C}{k} \sum_{\sigma(i) \in J_\omega} |Y_{\sigma(i)}(\omega) - \bar{Y}_{\sigma(i)}(\omega)| \\ &\to 0 \qquad \text{as} \quad n \to \infty.\end{aligned}$$

LEMMA 4. *Assume* (1.3) *and* (2.1). *Then*

$$\|V_n - \bar{V}_n\|_B \to 0 \qquad w.p.\ 1. \tag{3.3}$$

Proof.

$$\begin{aligned}\|V_n - \bar{V}_n\|_B &\leqslant \sum_{i=1}^{k} C_{ni} \|E[Y_{\sigma(i)} I(|Y_{\sigma(i)}| > n^{1/2}) \mid X_{\sigma(i)}]\|_B \\ &\leqslant \sum_{i=1}^{k} C_{ni} n^{-1/2} \|E[Y_{\sigma(i)}^2 \mid X_{\sigma(i)}]\|_B \\ &\leqslant C n^{-1/2}(D_2 + \|R\|_B^2) \to 0 \qquad \text{as} \quad n \to \infty.\end{aligned}$$

LEMMA 5. *Under the conditions of Theorem* 1, *it follows that*

$$\|\bar{T}_n - \bar{V}_n\|_B \to 0 \qquad w.p.\ 1. \tag{3.4}$$

The proof of Lemma 5 makes use of a standard inequality from [8].

LEMMA 6. *Let* $X_1, \ldots, X_n$ *be independent random variables satisfying* $|X_i| \leqslant M$, $EX_i = 0$ *and* $\operatorname{Var} X_i \leqslant \sigma^2$ *for all* i *and for some positive constants* M *and* σ^2. *Then, for* $0 \leqslant t \leqslant 2/M$,

$$E\left[\exp t\left(\sum_{i=1}^{n} X_i\right)\right] \leqslant \exp\left[nt^2\sigma^2 \frac{(1+tM)}{2}\right].$$

Proof of Lemma 5. Define $C(n,d)$ to be $2n^{d+1}$ if ρ is the Euclidean metric, or $2n^{2d}$ if ρ is the maximum component norm and $n > 2d$. It follows from the proof of Theorem 3 in [4] and the combinatorial inequalities in [14] that

$$
\begin{aligned}
&P[\sup_{x\in B} (\bar{T}_n(x) - \bar{V}_n(x)) > \varepsilon] \\
&\quad \leqslant C(n,d) \sup_{(x_1,\ldots,x_k)\in B^k} P\left\{ \sum_{i=1}^{k} C_{ni}[\bar{Y}_{\sigma(i)} - E(\bar{Y}_{\sigma(i)} \mid X_{\sigma(i)})] \right. \\
&\quad \left. > \varepsilon \mid X_{\sigma(i)} = x_i,\ i = 1,\ldots,k \right\} \\
&\quad \leqslant C(n,d)\, n^{-\varepsilon\beta_n} \sup_{(x_1,\ldots,x_k)\in B^k} E\left\{ \exp\left(\sum_{i=1}^{k} \beta_n \log nC_{ni}[\bar{Y}_{\sigma(i)} \right.\right. \\
&\quad \left.\left. - E(\bar{Y}_{\sigma(i)} \mid X_{\sigma(i)})] \right) \,\Big|\, X_{\sigma(i)} = x_i,\ i = 1,\ldots,k \right\} \\
&\quad \leqslant C(n,d)\, n^{-\varepsilon\beta_n} \sup_{(x_1,\ldots,x_k)\in B^k} \prod_{i=1}^{k} E\{\exp(\beta_n \log nC_{ni}[\bar{Y}_{\sigma(i)} \\
&\quad - E(\bar{Y}_{\sigma(i)} \mid X_{\sigma(i)})]) \mid X_{\sigma(i)} = x_i\}.
\end{aligned}
$$

There exists a sequence β_n such that $\beta_n \to \infty$, $\beta_n = o(\sqrt{n}/\log n)$, $k \geqslant [C\sqrt{n}\,\beta_n \log n]$. For each n fixed, let $M = 2n^{1/2}$ and $t_i = \beta_n \log nC_{ni}$, $i = 1,\ldots,k$. It is seen that $|\bar{Y}_{\sigma(i)} - E(\bar{Y}_{\sigma(i)} \mid X_{\sigma(i)})| \leqslant M$ w.p. 1 and $0 \leqslant Mt_i \leqslant 2n^{1/2}\beta_n \log nC_{ni} \leqslant 2$ by (1.3)(iii) and the choice of k. Further, $\sum_{i=1}^{k} t_i^2 \to 0$ as $n \to \infty$. From condition (2.1)(ii) and Lemma 6 we obtain that

$$
\begin{aligned}
&\sum_{n=1}^{\infty} P[\sup_{x\in B} (\bar{T}_n(x) - \bar{V}_n(x)) > \varepsilon] \\
&\qquad \leqslant \sum_{n=1}^{\infty} C(n,d)\, n^{-\varepsilon\beta_n} \exp\left| 3D_2 \left(\sum_{i=1}^{k} t_i^2 \right) \Big/ 2 \right| \\
&\qquad < \infty, \qquad \text{since } \beta_n \to \infty.
\end{aligned}
$$

The same argument implies that

$$
\sum_{n=1}^{\infty} P[\inf_{x\in B} (\bar{T}_n(x) - \bar{V}_n(x)) < -\varepsilon] < \infty.
$$

Now, (3.4) follows from the Borel–Cantelli lemma.

Proof of Theorem 1. Write the triangle inequality

$$
\| T_n - R \|_B \leqslant \| T_n - \bar{T}_n \|_B + \| \bar{T}_n - \bar{V}_n \|_B + \| V_n - \bar{V}_n \|_B + \| V_n - R \|_B .
$$

Theorem 1 now follows from (3.1) through (3.4).

Proof of Corollary 1. From Lemma 2 and Theorem 1, we have

$$\|\hat{T}_n - R\|_B \leqslant \|T_n(X_{\sigma(1)}) - R(X_{\sigma(1)})\|_B + \|R(X_{\sigma(1)}) - R(x)\|_B$$
$$\leqslant \|T_n - R\|_B + \|R_k\|_B \to 0 \qquad \text{w.p. 1.}$$

For the proof of Theorem 2, we shall provide rates of convergence for (3.1) to (3.4). All the terms defined in the proof of Theorem 1 except $\bar{Y}_j$ are retained. The only new definition is that for each positive integer n, $\bar{Y}_j = Y_j I[|Y_j| \leqslant n^{1/(2+d)}]$ for $j = 1,\ldots, n$. To establish Theorem 2 it remains to verify the following statements.

$$n^{1/(2+d)}\beta_n^{-1} \|V_n - R\|_A \to 0 \qquad \text{w.p. 1,} \tag{3.5}$$

$$n^{1/(2+d)} \|T_n - \bar{T}_n\|_B \to 0 \qquad \text{w.p. 1,} \tag{3.6}$$

$$n^{1/(2+d)}\beta_n^{-1} \|V_n - \bar{V}_n\|_B \to 0 \qquad \text{w.p. 1,} \tag{3.7}$$

and

$$n^{1/(2+d)}(\beta_n \log n)^{-1} \|\bar{T}_n - \bar{V}_n\|_B \to 0 \qquad \text{w.p. 1.} \tag{3.8}$$

We start with the proof of (3.5). Condition (2.2)(i) implies

$$\|V_n - R\|_A \leqslant \alpha \|R_k\|_A \qquad \text{w.p. 1,} \tag{3.9}$$

provided that $\|R_k\|_A \leqslant \delta$ w.p. 1. Thus, if it is shown that

$$n^{1/(2+d)}\beta_n^{-1} \|R_k\|_A \to 0 \qquad \text{w.p. 1,} \tag{3.10}$$

then (3.5) follows from (3.9) and (3.10). Assuming (2.2)(iv), Devroye and Wagner [5] showed that for any $\varepsilon > 0$

$$\sum_{n=1}^{\infty} P\left\{ \bigcup_{x: f(x) > \varepsilon} \left| v_d R_k^d > \frac{k}{n(f(x) - \varepsilon)} \right| \right\} < \infty \tag{3.11}$$

where v_d is the volume of the unit ball in R^d. Equation (3.11) implies that w.p. 1 for all but finitely many n,

$$R_k \leqslant (k/v_d n(f(x) - \varepsilon))^{1/d} \tag{3.12}$$

uniformly over any bounded set A where $\inf_{x \in A} f(x) > \varepsilon$. For $k = [Cn^{2/(2+d)}]$, (3.10) follows from (3.12) and (3.5) is obtained.

For proving (3.6), it is obvious that by replacing "$EY^2 < \infty$" with "$E|Y|^{2+d} < \infty$" the same argument of Lemma 3 establishes (3.6). Likewise, following the proof of Lemma 4, we can easily check that (3.7) is obtained from conditions (2.2)(ii) and (iii).

Finally, (3.8) will be derived via the same proof of Lemma 5. By setting

$M = 2n^{1/(2+d)}$ and $t_i = n^{1/(2+d)} C_{ni}$, $i = 1,..., k$, then applying Lemma 6, we notice that for any $\varepsilon > 0$

$$\sum_{n=1}^{\infty} P\{n^{1/(2+d)}(\beta_n \log n)^{-1} \| \bar{T}_n - \bar{V}_n \|_B > \varepsilon\} < \infty,$$

from which (3.8) follows. The proof of Theorem 2 is completed.

As a final remark, for the proof of Theorem 3, we note that (3.6) to (3.8) remain valid and the difference from the proof of Theorem 2 lies in obtaining a weaker relsult parallel to (3.5), i.e., for any $\varepsilon > 0$,

$$n^{(1-\varepsilon)/(2+d)} \| V_n - R \|_A \to 0 \qquad \text{w.p. 1.} \tag{3.13}$$

The proof of (3.13) (see [2]) utilizes a fundamental fact (see [9]) that the random variables $\mu(S(x; R_{\sigma(i)}))$, $i = 1,..., n$, are distributed like the order statistics from the uniform distribution over the interval [0, 1].

Acknowledgments

The author is grateful to Professor R. J. Serfling for helpful discussions and to a referee who pointed out an error in the early draft and made several comments leading to the present version.

References

[1] Cheng, K. F. (1979). Strong convergence in nonparametric estimation of a regression function. Statistics Report M510, Department of Statistics, The Florida State University.

[2] Cheng, P. E. (1981). *Strong Convergence of nearest Neighbor Regression Estimators.* Technical Report No. 347, *Department of Mathematical Sciences, The Johns Hopkins University.*

[3] Csörgö, M., and Révész, P. (1981). *Strong Approximations in Probability and Statistics.* Academic Press, New York.

[4] Devroye L. P. (1978). The uniform convergence of nearest neighbor regression function estimates and their application in optimization. *IEEE Trans. Inform. Theory* **IT-24** 142–151.

[5] Devroye, L. P., and Wagner, T. J. (1977). The strong uniform consistency of nearest neighbor density estimates. *Ann. Statist.* **5** 536–540.

[6] Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.* **9** 1310–1319.

[7] Fix, E., and Hodges, J. L., Jr. (1951). *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties.* Report No. 4, Project No. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

[8] Lamperti, J. (1966). *Probability.* Benjamin, New York.

[9] Mack, Y. P., and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *J. Multivariate Anal.* **9** 1–15.

[10] Royall, R. M. (1966). *A Class of Nonparametric Estimates of a Smooth Regression Function.* Ph.D. dissertation, Stanford University.

[11] SPIEGELMAN, C., AND SACKS, J. (1980). Consistent window estimation in nonparametric regression. *Ann. Statist.* **8** 240–246.

[12] STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.

[13] STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

[14] VAPNIK, V. N., AND CHERVONENKIS, A. YA. (1971). On the uniform convergence of the relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.